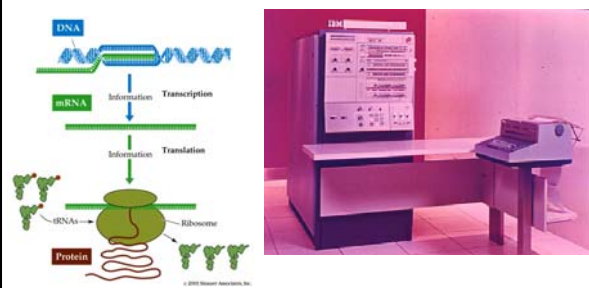


27622: Introduction to Bioinformatics, Turbo version

Henrik Nielsen, Associate professor
Center for Biological Sequence
Analysis

What is bioinformatics?



What are bioinformaticians up to, actually?

- *Manage* molecular biological data
 - Store in *databases*, organise, formalise, describe...
- *Compare* molecular biological data
- Find *patterns* in molecular biological data
 - *phylogenies*
 - *correlations* (sequence / structure / expression / function / disease)

Goals:

- *characterise* biological patterns & processes
- *predict* biological properties
 - low level data ⇒ high level properties
(eg., sequence ⇒ function)

CENTER FOR
BIOLOGICAL
CALCULATIONS
ENGINEERING
LYSIS CBS

Bioinformatics: neighbour disciplines

- Computational biology
 - Broader concept: includes computational ecology, physiology, neurology etc...
- -omics:
 - Genomics
 - Transcriptomics
 - Proteomics
- Systems biology
 - Putting it all together...
 - Building models, identify control & regulation

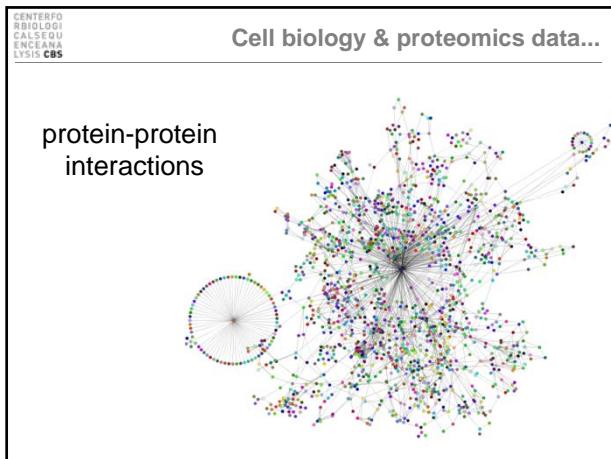
CENTER FOR
BIOLOGICAL
CALCULATIONS
ENGINEERING
LYSIS CBS

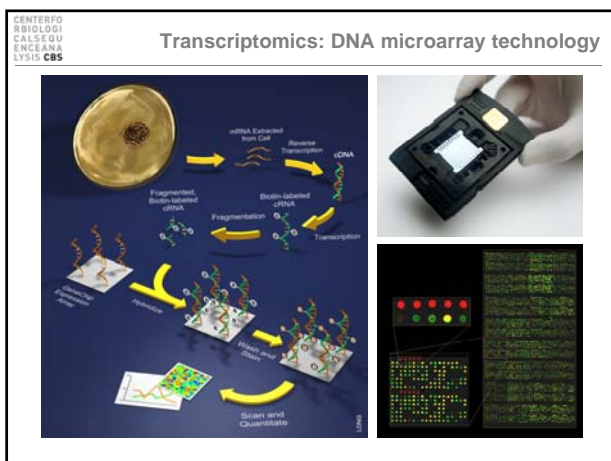
A view of Systems Biology

CENTER FOR
BIOLOGICAL
CALCULATIONS
ENGINEERING
LYSIS CBS

Bioinformatics: prerequisites

- **Bio- side:**
 - Molecular biology
 - Cell biology
 - Genetics
 - Evolutionary theory
- **-informatics side:**
 - Computer science
 - Statistics
 - Theoretical physics





CENTER FOR
RADIOLOGICAL
CALSEQU
ENCEANA
LYSIS CBS

Phenotype data: human diseases

CENTER FOR
BIOLOGICAL
CALSEQU
ENGINEERING
LYSIS CBS

Prediction methods

- **Homology / Alignment**
- Simple pattern ("word") recognition
- Statistical methods
 - Weight matrices: calculate amino acid *probabilities*
 - *Other examples*: Regression, variance analysis, clustering
- Machine learning
 - Like statistical methods, but parameters are estimated by iterative *training* rather than direct calculation
 - *Examples*: Neural Networks (**NN**), Hidden Markov Models (**HMM**), Support Vector Machines (**SVM**)
- Combinations

CENTER FOR
BIOLOGICAL
CALSEQU
ENGINEERING
LYSIS CBS

The computer

- *Everything* can be reduced to bits (0 or 1)



CENTER FOR
BIOLOGICAL
CALSEQU
ENGINEERING
LYSIS CBS

Digital information

- A byte = 8 bits

0 1 0 0 0 0 0 1

Can be interpreted as

- The number 65
- The letter "A"
- Part of a machine code instruction
- Part of a colour specification
- Part of a sound encoding
- ...

CENTERFO
R BIOLOGI
CAL SEQU
ENCE ANAL
YSIS CBS

Text files

A text file is a file where every byte is interpreted as a character

Examples

Plain text .txt
Program settings .ini
C source code .c
Python script .py
TeX source .tex
Web page source .html
Sequences .fasta

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	NUL	122	7A	Space	244	F4	À	266	FA	Ö
1	01	Start of heading	123	7B	!	245	F5	Á	267	FB	Ø
2	02	Start of text	124	7C	"	246	F6	Â	268	FD	Ù
3	03	End of text	125	7D	#	247	F7	Ã	269	FE	Ú
4	04	End of transmitted	126	7E	\$	248	F8	Ä	270	00	Û
5	05	Priority	127	7F	%	249	F9	Å	271	01	Ü
6	06	Address/length	128	80	&	250	FA	Ä	272	02	Ý
7	07	Audio link	129	81	'	251	FB	Å	273	03	Þ
8	08	End of sequence	130	82	(252	FC	Ä	274	04	ß
9	09	Non-printable	131	83)	253	FD	Å	275	05	ä
10	0A	Line feed	132	84	*	254	FE	Ä	276	06	å
11	0B	Carriage return	133	85	+	255	FF	Ä	277	07	æ
12	0C	Form feed	134	86	,	256	00	Ä	278	08	ç
13	0D	Carriage return	135	87	-	257	01	Ä	279	09	è
14	0E	Shift out	136	88	.	258	02	Ä	280	0A	é
15	0F	Shift in	137	89	/	259	03	Ä	281	0B	ê
16	10	End of message	138	8A	:	260	04	Ä	282	0C	ë
17	11	Device control 1	139	8B	;	261	05	Ä	283	0D	ì
18	12	Device control 2	140	8C	<	262	06	Ä	284	0E	í
19	13	Device control 3	141	8D	=	263	07	Ä	285	0F	î
20	14	Device control 4	142	8E	>	264	08	Ä	286	10	ï
21	15	End of transmission	143	8F	?	265	09	Ä	287	11	ï
22	16	End of transmission	144	90	@	266	0A	Ä	288	12	ï
23	17	End of transmission	145	91	A	267	0B	Ä	289	13	ï
24	18	Cancel	146	92	B	268	0C	Ä	290	14	ï
25	19	End of medium	147	93	C	269	0D	Ä	291	15	ï
26	1A	Substitute	148	94	D	270	0E	Ä	292	16	ï
27	1B	Ignore	149	95	E	271	0F	Ä	293	17	ï
28	1C	Not a separator	150	96	F	272	10	Ä	294	18	ï
29	1D	Not a separator	151	97	:	273	11	Ä	295	19	ï
30	1E	Not a separator	152	98	>	274	12	Ä	296	1A	ï
31	1F	Not a separator	153	99	?	275	13	Ä	297	1B	ï

The ASCII table

CENTERFO
R BIOLOGI
CAL SEQU
ENCE ANAL
YSIS CBS

Extended character sets

There are *many* ways to interpret characters with values above 127. Here, you see two of them.

"Mac OS Roman" Encoding

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
002	!	"	#	\$	%	&	'	()	*	+	,	-	.	/	
048	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
064	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
080	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
096	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
112	p	q	r	s	t	u	v	w	x	y	z	{		}	~	
128	À	Á	Â	Ã	Ä	Å	Ö	Ù	Ú	Û	Ü	Ý	Þ	ß	à	á
144	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó
160	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ß	à	á	â	ã
176	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó
192	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ß	à	á	â	ã
208	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó
224	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ß	à	á	â	ã
240	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï	ð	ñ	ò	ó

Windows-1252, sometimes called incorrectly "ANSI". Blue slots indicate unused or control characters.

CENTERFO
R BIOLOGI
CAL SEQU
ENCE ANAL
YSIS CBS

Text files—line endings

- UNIX standard (including Mac OS X):
 - 10 — LF ("Line feed" char).
- Old Mac (System 9 and before):
 - 13 — CR ("Carriage Return" char).
- DOS/Windows:
 - 13, 10 — both CR and LF.


A good text editor can handle all three systems.
Notepad for Windows *cannot*!

CENTER FOR
RADIOLOGICAL
CALCULATIONS
ENGINERING
LYSIS, CBS

www.jedit.org

jEdit
Programmer's Text Editor

Last Site Update: 19 November 2011 | Latest Version: 4 Spool 1 Stable Version: 4.4.2

 Download

jEdit is a mature programmer's text editor with hundreds (counting the more-developing plugins) of person-years of development behind it. To download, install, and set up jEdit as quickly and painlessly as possible, go to the [Quick Start](#) page.

While jEdit boasts many expensive development tools for features and ease of use, it is released as free software with full source code, provided under the terms of the [GPL 2.0](#).

The jEdit core, together with a large collection of plugins is maintained by a [world-wide developer team](#).

Some of jEdit's features include:

- Written in Java, so it runs on Mac OS X, OS/2, Unix, VMS and Windows
- Built-in macro language, extensible plugin architecture. Hundreds of macros and plugins available
- Plugins can be downloaded and installed from within jEdit using the "plugin manager" feature
- Auto-indent, and syntax highlighting for more than 200 languages
- Supports a large number of character encodings including UTF-8 and Thai-8
- Folding for selectively hiding regions of text
- Word wrap
- Highly configurable and customizable
- Every other feature, both basic and advanced, you would expect to find in a text editor. See the [Features](#) page for a full list.

Plugins

- Main Site
- Features
- Compatibility
- Screenshots
- News and Images
- Reviews
- Download
- Plugins

Community

- jEdit Community
- jEdit Wiki
- Quick Start Guide
- Online Documentation
- Feedback and Support
- Development
- SourceForge Project

sourceforge

JProfiler

Pygments
